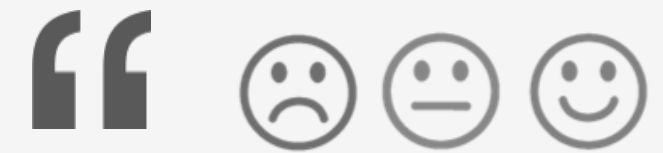


NLP 감정분석 기반
마케팅 시장 분석

온라인 커뮤니티 특화 감성 사전 구축을 위한 새로운 용어 극성값 분석 시스템

빅데이터미네이터



주제 수정: 온라인 커뮤니티 특화 감성 사전 구축을 위한 새로운 용어 극성값 분석 시스템

지난주 주요 피드백

- 기존 사전에 없는 새로운 단어에 대해 어떤 스탠스를 취할 것인가?
- 사용자 입장에서 모르는 단어를 받아들였을 때 어떻게 대응하는가?
- 기존에 알고 있는 단어와 어떻게 연관시켜서 추론하는가?
- 신조어의 패턴에 따라 극성값을 구분할 수 있는가?

구현 계획 구체화

1. 합성어, 커뮤니티 용어, 야민정음 등 **미리 KNU 사전 형태로 수동 입력**

→ 2명 이상 동의 하에 극성값 판단

→ KNU 사전의 신조어 및 인터넷 용어들이 너무 오래됨, 미리 입력하면 주변 단어들의 극성값이 더 정확하게 나올 것

2. 네이트판, FM에서 약 3일치 데이터 크롤링 후 **제목, 본문 합쳐서 형태소 csv 파일로 추출**

3. KNU 사전과 비교하여 **사전에 존재하지 않는 새로운 용어 모두 추출**

→ 빈도수에 따라 최종 "새로운 용어" 후보 결정

4. **FastText 기법**을 적용하여 새로운 용어의 **주변 단어 약 10개 명사 단위로 추출 후, 극성값 통계내서 출력**

→ **메인 기능** : 새로운 용어의 주변 단어 10개의 극성값을 통계 내서 제공함으로써,

사용자가 새로운 용어의 극성값을 판단하여 감성 사전에 추가해서 사용할 수 있도록 함

+ 추출한 용어들을 패턴에 따라 분류하고, 그 용어가 포함된 예시 문장을 제공하여 이해를 도움

→ 패턴 분류 기준 자료조사

→ 초성어, 합성어, 줄임말, 파생어로 카테고리 분류 가능

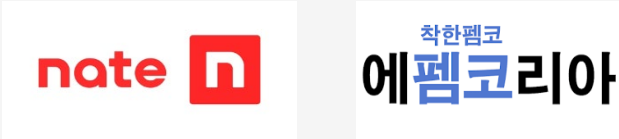
메인 기능 사용 기법 : FastText

- 페이스북에서 개발한 Word Embedding 기법
- 2020 이화여대 석사 논문 “한국어 텍스트에 대한 FastText 워드 임베딩 성능 연구”
 - Word2vec 보다 FastText를 사용할 때 감성분석 성과가 대체적으로 더 높게 나타남
 - FastText는 단어를 문자 단위 n-gram으로 나누고, n-gram 벡터들의 합으로 단어 벡터를 표현
 - 단어의 워드 임베딩이 가능할 뿐만 아니라, 임베딩 시 형태소적 특징을 반영
 - Word2vec에 비해 구문적(syntactic) 평가 측면에서 성능이 더 우수

표 7. 감성분석 모델 검증 결과

데이터	임베딩 모델	Accuracy	Precision	Recall	F1-score
데이터셋 1 (영화평)	Word2vec	0.6527	0.5143	0.7331	0.5899
	FastText	0.6560	0.5609	0.6942	0.6153
데이터셋 2 (뉴스 기사)	Word2vec	0.6674	0.6274	0.6836	0.6441
	FastText	0.7026	0.6723	0.6789	0.6693

진행 사항: 커뮤니티 크롤링&사전 구축



■ 크롤링: 네이트판&FM 코리아

1	제목 형태소	본문 형태소
2	[아기, '당', '호빵', '이]	['꼭', '잔', '개울', '을', '할', '해가는', '호빵', '이', '워낙', '운동량', '이', '적어서', '인자', '많아', '억자', '압아도', '통통', '냥', '보는', '사람', '을', '마
3	[활상, '친결하고', '진심', '으로', '대하', '러', '애쓰는데', '나', '는', '글', '홍자', '네]	[활상, '친결하러', '애쓰고잖', '대해', '주려', '애쓰고', '글', '진심', '인데', '활상', '홍자', '야', '나를', '파', '한다는', '것', '도', '느껴지고', '불편하
4	[사진, '은', '어떻게', '잘', '찍', '을', '수', '있는', '거니]	[아마, '이', '게', '실면서', '가장', '잘', '찍었다고', '생각', '하는', '사진', '임나', '도', '사진할', '꼭고시퍼', '왜', '계속', '찍어도', '그대로', '자', '그
5	[엄마, '의', '감정', '쓰레기통', '이었던', '나', '이런', '마음', '이', '흔한', '스러워]	[영할, '을', '앞두고', '친정', '에', '가야', '하나', '사회', '소년', '성일', '때', '첫째', '오빠', '집', '을', '사야하는데', '형약', '이', '없다며', '내', '영의
6	[자가, '격리', '중인', '시어머니', '께', '안부', '전화', '줄', '하러', '는', '남편]	[어머님, '이', '가게', '를', '하시는데', '꼭', '진자', '가', '다녀갔데요', '그래서', '어머님', '이', '검사', '후', '울성', '이산', '데주', '격리', '하계', '되
7	[다, '말했는데', '종아리', '만', '살아', '만', '빠지는', '이', '유가', '할', '까]	[다, '말했어요', '허리', '는', '인자', '말록', '전부', '오', '오죽', '말했으면', '들', '에', '책', '밖에', '없고', '허벅지', '도', '영영이', '도', '살', '없어서
8	[사소하지만, '엄마', '가', '행복해', '잘', '만', '한', '일이', '뭐', '가', '있을까]	[오죽, '잘', '웃', '창겨주고', '방', '에만', '있는', '거', '알아서]
9	[영할, '에는', '무조건', '시댁', '먼저', '가는', '게', '여의', '이고', '드리', '인가요]	[안녕하세요, '얼마', '전', '결혼', '한', '새댁', '입니다', '이번', '주석', '영할', '을', '앞두고', '또', '일이', '상졌네요', '처가', '댁', '시댁', '저희', '오
10	[전, '여자', '친구', '의', '물건', '을', '버리지', '않는', '남친', '의', '심리]	[사관지는, '년', '정도', '커를', '임', '중간', '에', '몇번', '헤어지고', '헤어짐', '중', '에', '한번', '은', '질게', '년', '헤어졌었을', '그', '사이', '동안', '가
11	[다, '들', '첫', '직장', '에서', '받은', '첫', '성과급', '어디', '에', '쓰셨나요]	[첫, '직장', '첫', '일사', '한', '회사', '에서', '첫', '성과급', '을', '받았어요', '어마', '무사', '하', '게', '큰', '돈', '은', '아니지만', '의미', '있게', '쓰고
12	[돌아가신, '할머니', '가', '너무', '보고', '싶은데', '시간', '이', '약', '할', '까요]	[제, '가', '속마음', '이나', '깊은', '애기', '를', '말', '놓고', '털어놓을', '데', '가', '없어서', '홍자', '삭히다가', '오늘', '은', '할머니', '가', '너무', '보
13	[후석, '전', '신랑', '과의', '언정', '누가', '더', '잘못', '할', '건가', '요]	[]
14	[나, '의', '동거', '모', '도', '절이', '분리', '불안]	[고양이, '도', '분리', '불안', '이', '있다고', '해서', '도', '절이', '는', '어떻게', '관할', '을', '해보았을', '고양이', '분리', '불안', '은', '질사', '가', '외
15	[살, '인데', '계대로', '될', '친구', '가', '한', '영도', '없습니다]	[오늘, '의', '편', '가지', '갈', '줄', '물랐', '네', '여', '외적인', '모습', '을', '가꾸라는', '분', '을', '이', '보이는데', '물줄', '이나', '물매', '다', '외적
16	[직장, '사수', '에게', '차려', '야', '하는', '신입', '의', '여의', '뭐', '가', '있나요]	[]
17	[살, '에', '신입', '으로', '취직', '이', '가능할는지', '모르겠습니다]	[살, '홍자', '독거', '노인', '으로', '살', '고', '있는', '남자', '입니다', '저', '는', '실업계', '를', '졸업', '하고', '군대', '전역', '후', '살', '부터', '바로', '가

▲	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	제목 형태: 본문 형태소														
2	[작, '은', '[숙담, '에', '나', '이런', '말', '이', '있', '다, 작은 고추, '가, '엎, '다, '보이, '는, '게, '다, '가, '아니, '라는, '속뜻', '을, '가지, '나, '숙담', '으로, '오늘, '소개, ']														
3	[아들놈, '[아들놈, '이, '를, '을, '시작, '하, '있, '는데, '못하, '게, '중, '막, '을, '수, '없, '나, '아름, '놈, '이, '욕, '먹, '는, '건, '상관, '없, '는데, '내, '명예, '를, '지켜														
4	[남, '고, ']														
5	[북한, '면 [미리, '가, '있, '어서, '유머, '탐]														
6	[현재, '미 [미국, '가, '마, '국경, '순찰대, '가, '말, '고백, '를, '채찍, '처럼, '휘두르, '며, 'Wu200b, 'Wu200b, '아이티, '난민, '을, '가죽, '물이, '하, '듯, '꽃이내, '는														
7	[경력, '이 [이력서, '에, '경력, '한, '줄, '은, '심하, '있, '지]														
8	[민병현, '[타격, '컨디션, '안, '중, '고, '몸, '아프, '나, '상황, '에서, '선수, '본인, '이, '군, '보내, '여, '달, '라, '하, '아도, '주장, '이, '라, '박, '고, '쓰, '나, '새끼, '고														
9	[공식, '발 [지나, '나, '할, '뇌동맥류, '수술, '을, '받, '았, '던, '민병현, '은, '온퇴, '후, '치료, '에, '전념, '하, '르, '계획, '이, '다, '민병현, '은, '선수, '생활, '종반, '가														
10	[땡, '죽구 [개죽, '워, '땡, '도, '무승부, '하, '는, '땡, '이, '나 데]														
11	[노, '운현 [화들짝]														
12	[오빠, '여 [결혼, '안, '하, '았, '으면, '아직, '남남, '아니, '나 가]														
13	[여자, '나 [을]														
14	[좋아하, '[머리, '속, '에, '마구, '나, '가, '그득, '하, '구나]														
15	[형제자매 []														
16	[살, '부, '이름, '하나하나, '언급, '하, '는, '게, '진짜, '귀, '엎, '네, '여]														
17	[다시, '보 []														
18	[노총참, '[무한도전, '로고, '새, '길]														
19	[아, '가, '[재, '기억, '상, '으로, '나, '될 '웨인, '노래, '이, '있, '던, '것, '같, '은데, '도우지, '찾, '을, '수, '가, '없, '음, '도, '입부, '이간, '데, '진짜]														
20	[배우, '정 []														
21	[사라지, '[저번, '주, '에, '속, '초, '가, '있, '을, '때, '파도, '가, '너무, '세, '이, '고, '백사장, '이, '침식, '되, '여, '있, '다고, '생각, '하, '았, '는데, '강릉, '도, '진짜,														
22	[내, '여자 [결혼, '하, '기, '로, '하, '고, '동거, '중, '이, '나 데, '지금, '고향, '에, '친구, '들, '만나, '러, '내려, '여, '감, '갔, '다, '오, '면, '안방, '에, '존, '나, '이, '쁘														
23	[인터넷, '[명창, '에서, '속도, '삭제]														
24	[안녕하세요 [오늘, '이, '재, '자, '살, '르, '생일, '이, '라, '축하, '의, '의미, '로, '올라, '여, '보, '보니다, '눈팅만, '하, '는, '아, '재안, '데, '중, '속스럽, '네요, '차량, '고														
25	[현재, '한 []														
26	[자기, '만 [이번, '글, '의, '주인공, '은, '자살, '소년, '고, '첫, '번, '째, '결혼, '고, '행복, '홍보, '창, '초기, '에, '는, '자살, '소년, '그대로, '올라가, '있, '지만, '자살, '가														

■ KNU 감성 사전에 기반한 커뮤니티 특화 감성 사전 구축

단어	뜻								
싫테	"싫어요(유튜브 기능) 테러"의 줄임말								
설참	"설명참고"의 줄임말								
닉차	"닉네임 차별"의 줄임말(닉네임을 보고 누군지를 판단해서 이를 근거해 차별하는 것을 의미)								
실매	실시간 매니저의 줄임말								
불소	"불타는 소통"의 줄임말								
홀로	올로+나홀로의 합성어로 본인 중심의 소비와 행복을 우선시하는 것								
시*비용	비속어인 시*+비용을 합친 단어로 스트레스와 같은 원인이 없었으면 쓰지 않았을 비용을 의미								
Z세대	밀레니엄 세대(1980년대 중반~1990년대 중반 사이)이후 태어난 세대를 일컫는 말								
가취관	같은 취향으로 모이되, 가벼운 모임을 선호하는 밀레니엄-Z세대								
마이싸이더	내안에 기준을 세우고 따르는 밀레니엄-Z세대								

싫테	-1	에펠탈 효과	+1
설참	0	윤족	0
닉차	-1	블루칼라	-1
실매	0	화이트칼라	+1
불소	+1	퍼플칼라	0
홀로	+1	미닝아웃	+1
시발비용	-1	미션 크리티컬	-2
Z세대	0	보이지 않는 고릴라	-1
가취관	+1	노컬 크러시	0
마이싸이더	0	스세권	+1
실감세대	+1	편세권	+1
유니콘 기업	+2	슬세권	+1
신파일러	-1	탈골공원	+1
82피플	-1	호캉스	+2
그린러쉬	0	나심비	+2

진행 사항: 커뮤니티 용어 패턴 조사

1. 줄임말: 실제 생활에서 자주 쓰여 줄이는 경우

Ex) 얼죽아, 만반잘부, 갑분싸 버카충, 지못미, 마상, 노잼, 싸강, 라방, 영끌, 주린이 등

2. 합성어: 두 단어의 합성어가 신조어가 되는 경우 ⇒ 비슷한 신조어의 생성이 계속해서 가능함

Ex) 배린이 → 방린이, 편리미엄 → 뷰리미엄, 학세권 → 스세권 등

3. 파생어: 접미사, 접두사 등 접사 파생으로 형성되는 신조어

Ex) -각, -충, -빠, -몬, -탱, 개-, 꿀-, 험- 등

4. 초성어

Ex) ㅇㅈ(인정), ㄱㅇㄷ(개이득), ㄹㄷ(레디), ㅇㄱㄹ(어그로) 등

5. 형태 변형 (야민정음)

Ex) 커엽다(귀엽다), 머통령(대통령), 머박(대박), 핍작(명작) 등

기대 효과



- 기존 감성 사전에 보유하고 있지 않은 새로운 용어에 대하여, FastText 기법으로 추출한 주변 단어 10개의 극성값을 통계 내어 제공한다. 사용자는 주변 단어 10개의 극성값 통계를 보고 새로운 용어의 극성값을 판단할 수 있다.
- Voting 방식으로 신조어의 극성을 추가하던 수동적인 방식에서 좀 더 자동화된 사전 구축 방식을 기대할 수 있다.
- 기존 온라인 커뮤니티 감성 분석에서 독립으로 추출되던 새로운 용어의 극성값을 파악할 수 있기 때문에, 사용자는 한층 더 품질 좋은 온라인 커뮤니티 특화 감성 사전을 구축하여 정확도가 향상된 감성 분석 결과를 얻을 수 있을 것으로 기대된다.

차주 계획

1. 커뮤니티 특화 감성 사전 지속적으로 업데이트
2. 크롤링 과정에서 전처리 부분 개선
3. 사전에 없는 새로운 용어 추출 모델 조사 및 Test



감사합니다 :)
빅데이터미네이터